A Solution to Metadata: Using XML Transformations to Automate Metadata

Jacqueline Mize
Radiance Technologies, Inc.
NOAA National Coastal Data Development Center
Building 1100, Room 101
Stennis Space Center, MS 39529

Christopher F. Robertson
Louisiana Office of Coastal Protection and Restoration
450 Laurel Street, Suite 1200
Baton Rouge, LA 70801

***Abstract*–There are many obstacles to creating and maintaining quality metadata. Producing and validating records against a particular standard can be both challenging and time consuming. More often than not, manually created records contain omissions and errors caused by poor record management tools and inadequate quality control measures. Some metadata standards are quite complex. In addition, it is clear that any manually executed generation, quality control, and management of metadata can be a resource drain on any organization. Therefore, *metadata automation*, which is the programmatic process of creating and updating metadata, is clearly the key to providing accurate metadata while also addressing the various challenges that any organization faces in balancing data stewardship needs with fiscal realities.**

Effectively using XML techniques and methods can achieve accurate, cost-efficient metadata. Overall, automation of metadata, using XML technologies, proved successful and provided many benefits as demonstrated through NOAA National Coastal Data Development Center's (NCDDC) partnership with the Louisiana Office of Coastal Protection and Restoration (OCPR). OCPR records resulted in the automated generation of 13,565 Federal Geographic Data Committee (FGDC) Content Standard for Digital Geospatial Metadata (CSDGM) compliant metadata records that were produced relatively quickly with few resources. Conventional methods of creating this metadata, using current metadata editing tools and template techniques, would have taken much longer and would have required significant additional resources. Thus, the automation has succeeded in producing far better results in terms of resources and time while increasing productivity.

I. Introduction

In 1990, the Office of Management and Budget (OMB) Circular A-16 called for Federal Agencies to create and maintain metadata, in accordance with the Federal Geographic Data Committee (FGDC) standards, for any spatial data that is collected, produced, acquired, maintained, distributed, used, or preserved [1]. President Bill Clinton signed Executive Order 12906 in 1994 [2] to strengthen OMB Circular A-16, which was later revised.

To protect the initial investment made in collecting data, an organization must commit resources to maintaining the official metadata record(s) for their data. Falling short of this, an organization runs the risk of degrading future potential data sharing and discovery and access.

Any organization involved in data collection must carry out their due diligence with regard to addressing data stewardship. This ultimately involves the generation of metadata to serve as the official record for the data as mandated by EO 12906. There are many obstacles to creating and maintaining quality metadata. Producing and validating records against a particular standard can be both challenging and time consuming. More often than not, manually created records contain omissions and errors caused by poor record management tools and inadequate quality control measures. Some metadata standards—such as the FGDC Content Standard for Digital Geospatial Metadata (CSDGM) [3] —are quite complex, necessitating the use of dwindling resources to train personnel in their proper use.

Metadata, the standardized documentation of data, comes in a variety of standards apart from the FGDC standards. Some standards predate the widely used FGDC CSDGM, and others were created to meet the specific needs of particular audiences. Many discipline-specific user communities, especially from the private and academic sectors, developed their own metadata standards—Directory Interchange Format (DIF), Ecological Metadata Language (EML), and International Organization for Standardization (ISO), to name a few. Metadata creation often is time consuming because many metadata standards are complex and difficult to implement. This variety of available standards has created some interoperability and compatibility issues. Many of the conventional metadata creation and validation methods in use today do not readily address interoperability issues.

Some core issues that must be considered include interoperability between systems and user communities, and compatibility among different metadata standards. Organizations may need to distribute metadata in a variety of

0-933957-38-1 ©2009 MTS

formats and standards to a diverse array of systems. For instance, an organization may have meticulously documented all their data using the EML standard. However, when later attempting to publish to a data clearinghouse—such as the Geospatial One-Stop (GOS), which requires that metadata be submitted using the FGDC or ISO standards [4]—the organization would find that it needs to adapt its metadata to satisfy this different standard without posing a further burden on organizational resources.

The solutions to metadata issues often lead to complex processes that become unsupportable. One such solution, crosswalks between standards, may address interoperability issues but usually results in the manual mapping of elements of one metadata standard to equivalent elements of another standard [5], usually within a spreadsheet application. Nonmappable elements are often left out, leading to a loss of information, or elements are mapped to nonequivalent elements and substandard metadata records are generated.

Using Extensible Markup Language (XML) techniques to automate metadata creation provides a way to overcome numerous obstacles to producing and maintaining relevant metadata. Programmatic metadata generation provides many other benefits, such as reduced effort, enhanced accuracy, and improved efficiency.

II. Automating Metadata

The Louisiana Office of Coastal Protection and Restoration (OCPR) collects large amounts of observed data to support ecological, hydrological, and climatological activities as part of their overall effort to evaluate the effectiveness of coastal restoration projects and coastal protection projects. The variety of coastal observations includes water quality, fisheries, and habitat data that are distributed from the Louisiana Department of Natural Resources (LDNR) web-accessible Strategic Online Natural Resources Information System (SONRIS) [6]. Mandated requirements for standardized metadata to accompany distributed data and to aid data management posed a significant resource challenge to OCPR. NOAA's National Coastal Data Development Center (NCDDC) partnered with OCPR to tackle metadata issues with an automated approach.

OCPR faced many familiar metadata roadblocks as they evaluated the requirements for creating standardized metadata to accompany their distributed data. To meet mandates and management requirements, OCPR needed FGDC CSDGM FGDC-STD-001-1998 compliant metadata for each coastal dataset distributed. Due to the large volume of coastal data distributed from LA DNR's 2008 SONRIS database, OCPR looked for more efficient and cost effective ways to meet these requirements. The NCDDC staff met with OCPR's Applied Coastal Engineering & Science (LACES) Division. The NCDDC Metadata Team was able to evaluate OCPR's metadata needs and partnered with OCPR to help resolve their metadata issues.

Metadata, when created in compliance with a standard such as FGDC, are a compilation of information about a data set in a particular format. The majority of elements required for various metadata standards often already exists within databases or is digitally documented from other sources, such as data dictionaries or standard operating procedures. This was the case for OCPR's coastal observational data. If metadata information is already digitally stored, it can be pulled from these sources to populate the metadata record, avoiding duplication issues. Metadata creation in these cases would require transferring the information provided from the databases and other sources to the targeted metadata standard elements. Format conversions from the source to the target format may be necessary. It was theorized that a programmatic process would be the ideal method of creating metadata, particularly for large data collections such as this particular collection of web-accessible coastal data.

Metadata automation, the programmatic process of creating and updating metadata, is the key to providing accurate metadata while addressing the various challenges that an organization faces in balancing data stewardship needs with fiscal realities. Metadata automation was the solution for the metadata problems that OCPR was facing.

III. Methods

Automated metadata can be generated by using XML techniques. A representative document of what a source contains, i.e., a data model, can be mapped to a representative document of the desired output, or target. These representative documents, called *schemas* (.xsd), can be created from XML, as subsequently described in detail. This mapping between the source and the target defines a *transform* (.xslt). The transform is then applied to the source XML to create the desired output. Fig. 1 is an overview of the process developed by the NCDDC using XML techniques to automate metadata creation.

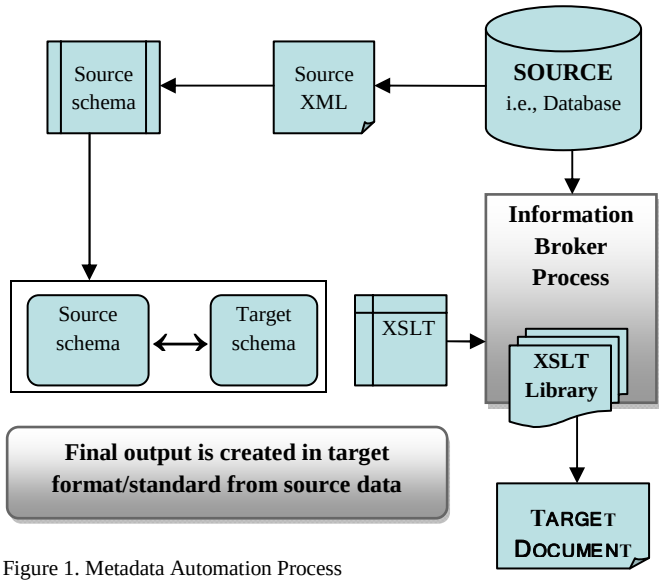

Figure 1. Metadata Automation Process

The use of XML technologies facilitates the exchange of complex structured data between applications. It protects the investment made in collecting the data by being flexible and easily adapted to technological changes, as well as being interoperable across the spectrum of formats and platforms [7]. It provides a rich set of standards that collectively support the automation of metadata, resulting in discrete process-steps that can be easily replicated and hosted on any platform. Additionally, XML tools are widely available that make the XML technology accessible to a broad range of implementers.

Microsoft Access® was used to develop the XML representation of the database. Database contents can either be directly imported into Access or be exported from the data source by creating a link from Access. Next, a query was created that was representative of all the data attributes. For example, if the data has collection stations and observation values in one table and coordinates for the stations in another table, the two tables must be joined to generate an output that includes all the properties for the data. The EXPORT function from the FILE menu in Access can be used to create the XML, which is then used to create the source XML schema definition (XSD) files for the data. Once created, these files can then mapped to corresponding metadata attributes of the target schema, such as was done for the OCPR metadata creation process.

Schemas of the data models are used to map information from other metadata standards from either a metadata-like database or other types of databases. Schemas of the XMLs are needed to define the structure, content, and semantics of the XML documents. The schema represents the data's model and defines the objects, attributes, and relationships while defining the rules for the structure and content of the XML document [8]. The proliferation of metadata schemas has provided a wide range to choose from as different communities attempt to meet the specific needs of users [5]. The existing FGDC-STD-001-1998 schemas [9, 10] were selected for use as the targeted format. Schemas can also be created from an existing XML document if there is a need for creation or customization. Because OCPR did not have any schemas of their databases, source schemas needed to be developed.

The development of schemas was accomplished using the XMLSpy® XML editor. XMLSpy® supports many features, including the ability for a user to create a schema based upon connections to external relational databases. XMLSpy® supports several of the most popular relational databases including Microsoft Access, Microsoft Structured Query Language (SQL) Server, Oracle®, ActiveX® Data Objects (ADO) compatible, and some Open Database Connectivity (ODBC) databases [11]. If no connection is possible, as was the case with OCPR's data, sample XML documents can be loaded into the editor and a schema can be generated [12].

The database schema, known as the source schema, can be mapped to the desired target schema. Mapping of schemas was accomplished with the use of XMLSpy's companion software, a visual programming tool called Altova Map-Force® [12]. MapForce is an essential integration tool for XML and database development that requires little or no programming knowledge and skill; however, a working knowledge of schemas and metadata standards is recommended.

Once the source schema and the target schema are selected, the mapping process can begin. Elements between the source and target that crosswalk can be directly connected. The dynamic elements of the OCPR data, which did not require conversions, were mapped from the source to the target in this manner.

MapForce contains multiple libraries with individual functions. Depending on the desired programming language output, supporting functions appear as boxes that can be simply selected and dragged into place within the mapping. Occasionally, the source schema does not contain all of the data needed to comply with the target schema. Constants may be required to create boilerplates for persistent information, which are practical and convenient for static elements. Boilerplates are also useful for providing missing information for elements that do not map.

For the OCPR schema that was to be created, each record generated would have the same Metadata Reference Information. As seen in Fig. 2, constants were added for this persistent information. These hardcoded elements completed the entire mandatory FGDC Metadata Reference Information section [3]. Methods of collecting and analyzing the data, as well as quality assurance and quality control methods, were fully documented and accessible online [13]. Because these process steps and QA/QC processes were already documented and available, static links were also hardcoded into the transform.

Some elements may have one-to-one relationships, but many do not. Simple functions may be required to produce the desired outcome. Concat is an example of a simple function that combines two or more elements from the source schema or constants and places the result in a single element in the target document as shown in Fig. 3 [11]. The reverse process may also be needed. Single elements from the source can be divided through various string and logic functions and mapped to one or more target elements.

Selective elements from the source schema may need to be interpreted and translated to conform to specified standards of the target schema. The user has the ability to select a field from the source schema, instruct how to interpret the input, and add the resulting interpretation(s) to the target schema.

Some mappings may require additional process steps that are not supported by the default library of functions. MapForce supports the creation of user-defined functions to address this need. Once defined, these new functions are available in the same manner as the default functions [14].

To address the complexity of large process chains, user-defined functions can be developed which encapsulate several steps, thus reducing the visual clutter in the interface and improving readability of mapping details.

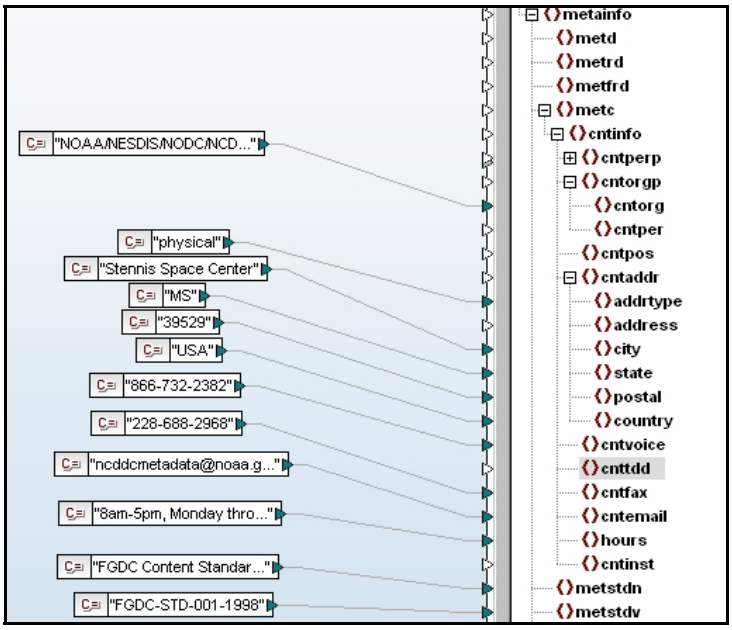

Figure 2. Constants added to mapping (from Ref. 11).

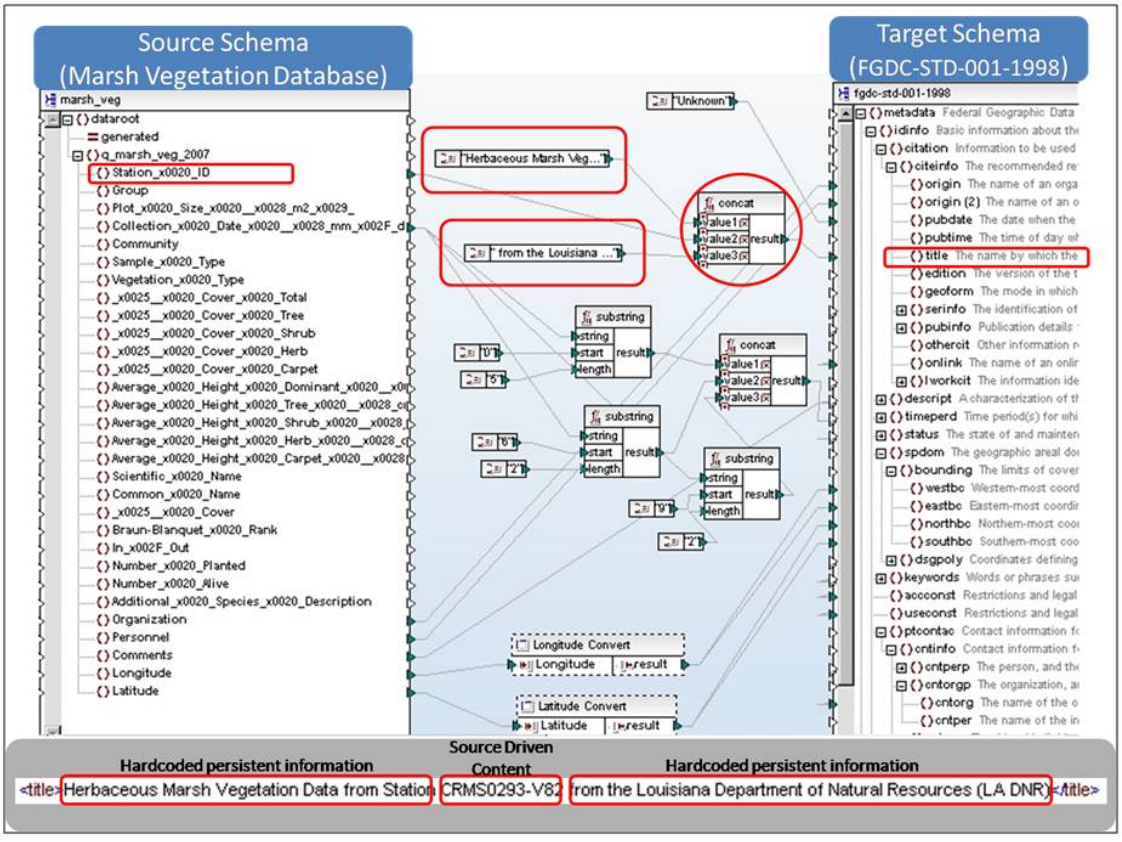

Figure 3. Concat function joining persistent information with source-driven content (from Ref. 11).

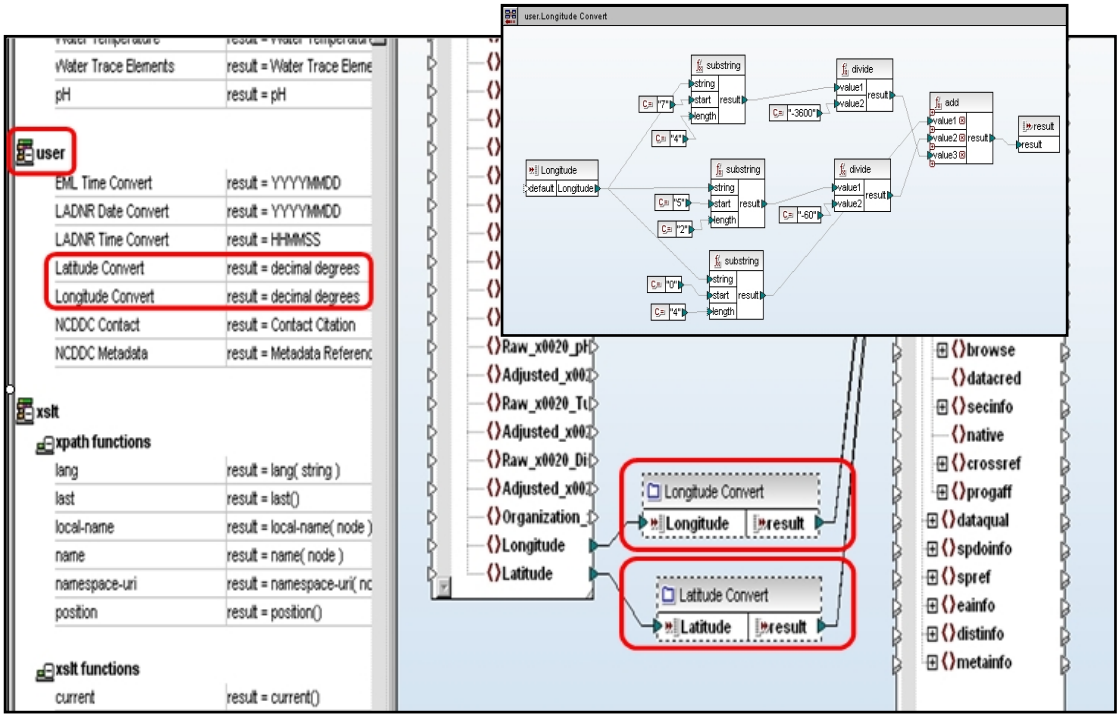

Figure 4. User-defined functions show conversion of latitude and longitude (from Ref. 11).

```
<?xml version="1.0" encoding="UTF-8"?>
<!--
This file was generated by Altova MapForce 2008sp1

YOU SHOULD NOT MODIFY THIS FILE, BECAUSE IT WILL BE
OVERWRITTEN WHEN YOU RE-RUN CODE GENERATION.

Refer to the Altova MapForce Documentation for further details.
http://www.altova.com/mapforce
-->
<xsl:stylesheet version="1.0" xmlns:xsl="http://www.w3.org/1999/XSL/Transform" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:xs="http://www.w3.org/2001/XMLSchema" exclude-result-prefixes="xs xsi xsl">
  <xsl:output method="xml" encoding="UTF-8" indent="yes"/>
  <xsl:template match="/">
    <metadata>
      <xsl:attribute name="xsi:noNamespaceSchemaLocation">
        <xsl:value-of select="'Q:/users/mize.jacqueline/mapping/schemas/MERMAI~1.0/CSDGM/fgdc-std-001-1998.xsd'"/>
      </xsl:attribute>
      <xsl:variable name="var1_instance" select="."/>
      <idinfo>
        <citation>
          <citeinfo>
            <xsl:for-each select="$var1_instance/dataroot/q_marsh_veg_2007/Organization">
              <origin>
                <xsl:value-of select="string(.)"/>
              </origin>
            </xsl:for-each>
            <xsl:for-each select="$var1_instance/dataroot/q_marsh_veg_2007/Personnel">
              <origin>
                <xsl:value-of select="string(.)"/>
              </origin>
            </xsl:for-each>
            <pubdate>
              <xsl:value-of select="'Unknown'"/>
            </pubdate>
            <xsl:for-each select="$var1_instance/dataroot/q_marsh_veg_2007/Station_x0020_ID">
              <title>
                <xsl:value-of select="concat(concat('Herbaceous Marsh Vegetation Data from Station ', string(.)), ' from the Louisiana Department of Natural Resources (LA DNR)')"/>
              </title>
            </xsl:for-each>
```

Figure 5. XSLT generated from MapForce (from Ref. 11).

Complex mathematical functions and recurring translations are prime candidates that can benefit from creating user-defined functions. For example, OCPR collects longitude and latitude data in degree–minute–second format**,** but FGDC metadata standards require the longitude and latitude to be in decimal–degree format. Instead of re-creating this complex mathematical function each time the coordinates needed to be converted, this conversion was created as a function within a user-defined library. Fig. 4 shows this commonly occurring conversion of latitude and longitude from degrees, minutes, and seconds to decimal degrees. This process provides a simple drag-and-drop function analogous to the MapForce default libraries for the selected output language.

Throughout the mapping process, MapForce checks the validation of the mappings against the assigned target schema. If errors exist [14], the generation of an Extensible Stylesheet Language Transformation (XSLT) will abort the mapping process. MapForce also allows a sample XML record to be applied to the source. Throughout the mapping process, the transform can be monitored using a preview of the mapping result of the applied sample XML record for expected results.

Once the mapping is complete and valid, the visual programming tool builds the transform (XSLT), which can support a variety of languages. MapForce supports the creation of an XSLT in XSLT (xpath1.0), XSLT2(xpath2.0), XQuery, Java, C#, and C++ [12]. The generated XSLT can be used either "as is" or further edited in an XML editor. Output from XSLTs can be XML, HTML, or plain-text documents as well. OCPR transforms were generated as xpath 1.0, seen in Fig. 5.

Once the XLS transform is complete, it can be added to NCDDC's Information Broker Service transform library. The Information Broker uses a third-party XSLT engine to perform XSL transformations. Clients of the Information Broker construct calls to its transform(s) method by providing the type of the incoming content (e.g., eml), the type of the resulting content (e.g., fgdc) and the XML content to be transformed. The result of this service call is the transformed content.

Seven XSLT's were built, one for each of the coastal data types. Most of the information for the meta-data records was thoroughly documented in various other documents**,** and stationary links to these documents were added to boilerplate elements as required. The resulting XSLTs were added to the transform library within the Information Broker Service**,** resulting in new FGDC-compliant metadata records.

IV. Results

This mapping process was used to fully automate the creation of valid FGDC-compliant metadata for all OCPR coastal data, which resulted in 13,565 FGDC CSDGM metadata records. The resulting OCPR records contained much more information than the minimum requirement of complete "Identification Information" and "Metadata Reference Information" sections. The records also included complete "Data Quality Information," "Entity Attribute Information," and "Distribution Information" sections.

These resulting metadata records were subjected to QA/QC techniques. Random sampling was conducted on four percent of the resulting OCPR records to check for validation against the FGDC CSDGM schema using NCDDC's MERMAid tool [15]. All randomly sampled records passed validation. Record content was also visually inspected on randomly selected records by several OCPR and NCDDC staff.

Time spent mapping the schemas and creating the XSLTs took one metadata specialist about two weeks. The metadata generation process for OCPR records, on average, took approximately 0.9 second of real time per record. Only a few exceedingly large files within the Continuous Hydrographic and DCP database took much longer to process than the rest. These rare cases are considered outliers and are not included in the average processing.

IV. Conclusions

Automation of metadata using XML technologies proved to be successful and provided many benefits. Over 13,000 FGDC CSDGM compliant metadata records were produced quickly, dramatically reducing record management overhead for the organization.

Programmatic generation of metadata allowed for greater consistency among the records. The amount of errors and omissions can be limited by the automation process. All records processed using the same transform will be processed in the same consistent manner.

Record maintenance was effectively reduced to maintaining the accuracy of the resulting XSLTs. Updates can be applied in one location and applied to all records. As changes occur at databases, such as corrections, additions, or deletions of data, the entire record inventory can be reprocessed programmatically. Also, automated metadata can be scheduled to rerun periodically as the database is updated so that the metadata remains current and accurately reflects any changes in the data. Any changes that affect the existing accuracy of the transform(s) can be updated within the transform(s), and the entire record inventory can be reprocessed programmatically. The coastal data from OCRP has confirmed that current metadata can be produced and maintained in this manner with minimal resource usage.

Interoperability between systems, interoperability between user communities, and compatibility among different metadata standards can be made easier through automation of metadata. Automation makes the transition to other metadata standards, such as ISO and NAP, a manageable process. Multiple transforms can be applied to a source to create output in a variety of formats and standards. Granularity

of metadata records, either at collection level or at individual record level, can be addressed programmatically, depending on how the database is queried. The current process generates a metadata record for each row in the query output. If the number of records is too large, then the EXPORT tool in MS Access cannot handle the process, and the output has to be divided into multiple sets by changing the conditions in the query. For example, if the data includes multiple years, the query can be adjusted to create separate files for each year.

Direct connections to the databases and subsetting large files could improve efficiency and reduce processing time. No direct connections to the databases were established because of the precautionary measures taken for testing purposes. The process of creating the XML and XSD files for the data is simple, but can be time consuming because the EXPORT tool is extremely slow when dealing with a large number of records.

Adding conditions to the query for limiting the number of records to generate the metadata conversion map could possibly decrease the processing time. Exceedingly large records may also increase processing time.

The greatest potential for error occurs during the mapping process. The users' level of familiarity with the source data and the target schema greatly affect the accuracy of the resulting metadata. The success of this process hinges ultimately on the accuracy of the mapping effort. At this stage of the process, it is vital to utilize QA/QC techniques and review to achieve quality products.

V. REFERENCES

[1] Office of Management and Budget, Coordination of Geographic Information and Related Spatial Data Activities. OMB Circular A-16, Rev. August 19, 2002.

[2] President Bill Clinton, United States Executive Order 12906, April 11, 1994. Coordinating Geographic Data Acquisition and Access: *Federal Register* 59(71).

[3] Federal Geographic Data Committee, FGDC-STD-001-1998, Content Standard for Digital Geospatial Metadata, Washington, D.C., rev. June 1998.

[4] Geospatial One-Stop, Creating and Publishing Metadata in Support of the Geospatial One-Stop and the NSDI, *Geospatial One-Stop*, geodata.gov/gos/metadata/CreatePublishMetadata.pdf, July 31, 2006; accessed January 6, 2009.

[5] Lois Mai Chan and Marcia Lei Zeng, "Metadata Interoperability and Standardization—A Study of Methodology, Part I. Achieving Interoperability at the Schema Level," *D-Lib Magazine* 12(6).

[6] Louisiana Department of Natural Resources, "SONRIS," sonris-www.dnr.state.la.us/www_root/sonris_portal_1.htm; accessed Dec. 22, 2008.

[7] Ed Tittel, Natanya Pitts, and Frank Boumphrey, *XML for Dummies*, 3rd ed. New York: Hungry Minds, Inc., 2002.

[8] Altova, *XMLSpy 2004, Enterprise Edition: User & Reference Manual*. Beverley, MA: Altova GmbH & Altova, Inc., 2003.

[9] Richard E. Rathman, Mike Moeller, and Doug Nebert, *FGDC Metadata XML Schema 1.0.0 20030801*, USGS/NOAA Coastal Services Center, Federal Geographic Data Committee, Washington, DC, 2003.

[10] NOAA Coastal Services Center, CSDGM XML Schema Document Representation, fgdc.gov/metadata/fgdc-std-001-1998.xsd, accessed December 22, 2008.

[11] Altova, MapForce (computer software), Professional Edition, version 2008 sp1. Beverley, MA: Altova, 2008.

[12] Altova, XMLSpy (computer software), Professional Edition, version 2008 sp1. Beverley, MA: Altova, 2008.

[13] Louisiana Department of Natural Resources, Coastal Restoration Division, "Louisiana Department of Natural Resources, Strategic Online Natural Resources Information System, SONRIS 2000, Coastal Restoration Division Biological Database, Data Descriptions," dnr.louisiana.gov/crm/coastres/projectdata/DataDescriptions.pdf, 2008, accessed July 9, 2009.

[14] Altova, *MapForce 2005, User and Reference Manual*. Beverley, MA: Altova GmbH & Altova, Inc., 2003.

[15] NOAA National Coastal Data Development Center, MERMAid (computer software), version 1.2, Stennis Space Center, MS.